

On the estimation of the order of smoothness of the regression function^{*}

Daniel Taylor-Rodriguez[†]

Sujit K. Ghosh[‡]

October 13, 2015

Abstract

The order of smoothness chosen in nonparametric estimation problems is critical. This choice balances the tradeoff between model parsimony and data overfitting. The most common approach used in this context is cross-validation. However, cross-validation is computationally time consuming and often precludes valid post-selection inference without further considerations. With this in mind, borrowing elements from the objective Bayesian variable selection literature, we propose an approach to select the degree of a polynomial basis. Although the method can be extended to most series-based smoothers, we focus on estimates arising from Bernstein polynomials for the regression function, using mixtures of g-priors on the model parameter space and a hierarchical specification for the priors on the order of smoothness. We prove the asymptotic predictive optimality for the method, and through simulation experiments, demonstrate that, compared to cross-validation, our approach is one or two orders of magnitude faster and yields comparable predictive accuracy. Moreover, our method provides simultaneous quantification of model uncertainty and parameter estimates. We illustrate the method with real applications for continuous and binary responses.

1 Introduction

Consider the following regression problem. Let $(X, Y) \in [0, 1] \times \mathbb{R}$ be a random vector such that $E[Y^2] < \infty$ and $E[Y|X = x] \equiv \mu(x)$, where $\mu(\cdot)$ is continuous on $[0, 1]$. For a random sample of n observations $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$, we assume the model is given by

$$Y_i = \mu(X_i) + \sigma\epsilon_i, \text{ for } i = 1, \dots, n,$$

^{*}This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[†]Postdoctoral Fellow, Duke University / SAMSI, RTP, NC 27709. Email: taylor-rodriguez@samsi.info

[‡]Professor, NCSU / Deputy Director, SAMSI, RTP, NC 27709. Email: ghosh@samsi.info

where $\sigma > 0$ and $\mu(\cdot)$ are parameters to be estimated, and ϵ is such that $E[\epsilon|x] = 0$ and $\text{Var}(\epsilon|x) \leq 1 \forall x \in [0, 1]$.

Establishing the functional relationship between the response Y and the predictor X has been a recurring theme in the statistical and mathematical literature for the past several decades. Parametric forms determine a priori the functional form of this relationship. As such, these are limited in their ability to uncover the true nature of the association. On the other hand, a plethora of nonparametric methods (see Ruppert et al., 2003, and references therein), which adapt to nonlinear features present in the data and enjoy good theoretical properties, have been extensively studied (e.g., Stadtmüller, 1986).

Nonparametric methods can be classified roughly into four categories: smoothing splines (Craven and Wahba, 1978; Eubank, 1985; Silverman, 1985), regression splines (Friedman, 1991; Wang, 2002; Wood, 2003), penalized splines (Eilers and Marx, 1996, 2010), and kernel-based methods (Nadaraya, 1964; Ramsay, 1991). The efficacy of nonparametric techniques in approximating $\mu(\cdot)$ is highly dependent on the choice of the *smoothness parameters* that regulate the balance between model generality and fidelity to the observed data. In principle, the nonparametric take on the problem described above rests on the fact that an unknown square integrable function $\mu(\cdot)$ can be represented using basis function expansions of the form

$$\mu(x) = \sum_{k=0}^{\infty} \beta_k \psi_k(x), \quad (1)$$

where the known functions $\psi_0(\cdot), \psi_1(\cdot), \psi_2(\cdot), \dots$ form a basis with respect to a suitably chosen metric norm. It is well-known a fact that any square integrable function can be approximated arbitrarily well by a smooth function supported on a compact set (e.g., see Appendix A in Györfi et al., 2002).

However, in practice, having an infinite or an excessively large number of terms in the sum leads to overfitting the model to the observed data. To prevent this, the smoothness of the function is controlled by selecting a suitable number of terms; we refer to this number of terms as the *order of smoothness*. Our goal is then to estimate a smooth function by determining where to truncate the infinite series, and provide valid inference about the order itself.

In this context, cross-validation is one of the most popular procedures used to select the order of smoothness (see Bouezmarni and Rolin, 2007; Leblanc, 2010). In spite of this, cross-validation is computationally costly (time-wise), and the inference conducted on the model parameters after cross-validation (and all other available methods used to select the order of smoothness, e.g., Manté, 2015), commonly ignores additional uncertainty arising from the selection procedure (see Berk et al., 2013, for a general description of this issue). However, a procedure to choose optimally and automatically the degree of the polynomial while accounting for uncertainty, has remained elusive.

As an alternative, in this article we introduce an approach that finds a suitable order of smoothness (hereafter denoted by \mathcal{N}^*), by approximately minimizing a predictive loss, while assessing the uncertainty associated with this choice. The proposed method is automatic, in the sense that no user input is required (as “noninformative” priors are used on model parameters), it is com-

putationally fast, and, can be extended to the case with multiple predictors and discrete-valued responses. As shown in Section 5, our method yields results equivalent to those obtained with cross-validation in terms of accuracy, while taking only a fraction of the time and selecting more parsimonious models than cross-validation.

Among the many nonparametric techniques available, our interest lies in the class of series-based smoothers, focusing on approximations to the unknown regression function $\mu(\cdot)$ with Bernstein polynomials (BPs). This choice facilitates the exposition of the procedure and benefits from the long list of virtues that BPs possess (see Farouki, 2012, and references therein).

It is well known that Bernstein bases span the space of continuous functions, $C[0, 1]$, but the elements of this basis are not orthogonal. Orthogonality of the basis is desirable as it guarantees *permanence* of the coefficients with respect to the degree of the approximation; meaning that, in approximations with orthogonal polynomials of orders K and K' (for $K \neq K'$ and $k \leq \min\{K, K'\}$), the order- k coefficients coincide. Selection of the order of smoothness cannot be performed directly on the Bernstein form; however, linear maps can be built between Bernstein polynomials and other orthogonal bases (see Farouki, 2012; Lee et al., 2002, and references therein). Moreover, it has been shown under orthogonality, using squared error loss, that the Bayesian median probability model (MPM) is optimal for prediction, provided that some additional conditions hold (see Barbieri and Berger, 2004). In Section 2, we provide some more details about BPs and their connection to other polynomial basis.

The existence of linear maps between Bernstein and orthogonal bases implies that choosing the order of smoothness in the space of orthogonal polynomials imposes an order onto the Bernstein polynomial. Hence, since each choice for the order of smoothness is associated to a specific model, we may cast this as a Bayesian variable selection problem (with an orthogonal base) where the polynomial hierarchy of term inclusion is respected (Chipman, 1996; Taylor-Rodriguez et al., 2015). This type of constrained selection implies that, for the k th degree basis to be included in the model, the inclusion of all coefficients of order $k - 1$ and lower is necessary. Conversely, if the k th degree basis is excluded from the model, this constraint prevents any term of order greater than k from being included.

The article is organized as follows. Section 2 provides details about BPs, emphasizing the connection to other orthogonal polynomial bases that will allow us to make use of variable selection tools. In Section 3 we introduce some underlying concepts on Bayesian selection, formulate the problem and elaborate on the proposed methodology. Section 4 provides some results demonstrating, under squared error loss, the asymptotic predictive optimality for the order of smoothness selected. In Section 5, through simulation experiments we assess the performance of our approach and compare it to that of cross-validation. Section 6 illustrates the use of the methodology with two case studies and describes a simple adaptation to incorporate binary responses. Finally, we close by providing some concluding remarks.

2 Preliminaries: Bernstein polynomials and orthogonal bases

Bernstein polynomial type estimators of regression functions were first developed by Stadtmüller (1986) for the single predictor case. Tenbusch (1997) extended the methodology for the multi-

ple predictor regression problem, and demonstrated that Bernstein polynomial type estimators are pointwise and uniformly consistent, and are asymptotically normal. These properties are also inherited by its derivatives. In addition to the above, the Bernstein basis is the only *optimally-stable* basis function in use (Farouki, 2012). This implies that, among all non-negative polynomial bases, its coefficients are the least vulnerable to the effects of random perturbations. Theoretical underpinnings aside, curve estimation with BPs is straightforward, can be used with censored data (Osman and Ghosh, 2012), and allows imposing shape constraints on the curve effortlessly whenever background information to do so is available (Curtis and Ghosh, 2011; Wang and Ghosh, 2012).

The BP approximation of degree $\mathcal{N} \in \mathbb{N} \cup \{0\}$ to the function $\mu(\cdot)$ at a value $x \in [0, 1]$ is given by a sequence of polynomials of the form

$$\begin{aligned}\mu_{\mathcal{N}}^{(\mathbf{B})}(x) &= \sum_{k=0}^{\mathcal{N}} \eta_k^{\mathcal{N}} b_k^{\mathcal{N}}(x) \\ &= (b_0^{\mathcal{N}}(x), \mathbf{b}_{\mathcal{N}}(x)')' \begin{pmatrix} \eta_0 \\ \boldsymbol{\eta}_{\mathcal{N}} \end{pmatrix},\end{aligned}\tag{2}$$

where one may choose $\eta_0^{\mathcal{N}} = \mu(0)$ and $b_0^{\mathcal{N}}(x) = (1-x)^{\mathcal{N}}$, and $\eta_k^{\mathcal{N}} = \mu(k/\mathcal{N})$ and $b_k^{\mathcal{N}}(x) = \binom{\mathcal{N}}{k} x^k (1-x)^{(\mathcal{N}-k)}$, for $k = 1, \dots, \mathcal{N}$. Other choices for $\eta_k^{\mathcal{N}}$ can be obtained using through iterated Bernstein polynomial operators (see Kelisky and Rivlin, 1967; Manté, 2015).

The choice of \mathcal{N} is critical in producing a good approximation as it is well know that, if $\mu(\cdot)$ satisfies an order $\alpha \in (0, 1]$ Lipschitz condition, then (Berens and Lorentz, 1972)

$$\|\mu(\cdot) - \mu_{\mathcal{N}}^{(\mathbf{B})}(\cdot)\|_{\infty} \equiv \sup_{0 \leq x \leq 1} |\mu(x) - \mu_{\mathcal{N}}^{(\mathbf{B})}(x)| = O(\mathcal{N}^{-\alpha/2}).$$

Asymptotically, Stadtmüller (1986) obtained as an optimal choice $\mathcal{N} = O(n^{2/5})$; similarly Tenbusch (1997) suggested using some integer $\mathcal{N} \in [n^{2/5}, n^{2/3}]$ (provided certain conditions hold, see Tenbusch (1997) for details).

Having these values to guide the selection of the order of the polynomial is useful, but once the sample size becomes moderately large, having the order of smoothness lie in $[n^{2/5}, n^{2/3}]$ can make the method susceptible to overfitting (see simulations in the Appendix). As such, an integer in the interval $[n^{2/5}, n^{2/3}]$ could instead be taken as an upper bound for the order of the Bernstein polynomial rather than being used as the order of the polynomial itself. Throughout the remainder of the article, we will take $\mathcal{N} = \lfloor n^{2/3} \rfloor$ as the upper bound on the order of smoothness.

In principle, the Bernstein degree- \mathcal{N} approximation given in (2) may be represented in terms of any degree- \mathcal{N} polynomial basis. Any such change of basis is obtained through a linear map, which produces the coefficients of the new basis by multiplying the coefficients from the original one by a transformation matrix $\mathbf{Q}_{\mathcal{N}}$ (Farouki, 2012).

The linear map between the coefficients of the Bernstein basis and those of an orthogonal basis (characterized by a transformation matrix $\mathbf{Q}_{\mathcal{N}}$) is extremely convenient for the problem at hand. One may first select the order of smoothness by choosing the polynomial model of order \mathcal{N}^* with the orthogonal basis, estimate its parameters, and finally, transform these estimates into their BP

form using the transformation matrix $\mathbf{Q}_{\mathcal{N}^*}$. Alternatively, to avoid the loss of stability that arises from going from one basis to another, one may simply use the selected order of smoothness \mathcal{N}^* under the orthogonal representation, and fit directly the order- \mathcal{N}^* BP to the data to obtain the parameter estimates.

2.1 Choice of orthogonal basis

The Bernstein basis is the only *optimally-stable* basis function in use (Farouki, 2012). Optimal stability implies that, among all non-negative polynomial-bases, its coefficients are least vulnerable to random perturbations. Consequently, it is usually recommended to avoid such transformations to prevent increased amplification of random perturbations in the coefficients. Transformation stability from the Bernstein to other bases has been studied extensively (see Farouki, 2012, and references therein), and is commonly assessed with what is referred to as the *condition number*. The *condition number* quantifies how much a random perturbation on an input (the coefficients) impacts the output (the curve estimate). The choice of orthogonal basis is all-important to control the loss of stability in the estimates. Among the transformations between Bernstein and other commonly used bases (e.g., power, Chebychev and Legendre), that onto Legendre polynomials has been shown to retain the most stability (Farouki, 2000).

The Legendre polynomial of degree \mathcal{N} in $[0, 1]$ is defined by the recurrence relation defined by

$$\begin{aligned}\psi_0(x) &= 1 \\ \psi_1(x) &= 2x - 1 \\ (k+1)\psi_{k+1}(x) &= (2k+1)(2x-1)\psi_k(x) - k\psi_{k-1}(x), \quad k = 1, 2, \dots\end{aligned}$$

The order \mathcal{N} approximant to $\mu(x)$ in Legendre form is given by (Farouki, 2012)

$$\begin{aligned}\mu_{\mathcal{N}}^{(\mathbf{L})}(x) &= \sum_{k=0}^{\mathcal{N}} \lambda_k \psi_k(x) \\ &= (\psi_0(x), \boldsymbol{\psi}_{\mathcal{N}}(x))' \begin{pmatrix} \lambda_0 \\ \boldsymbol{\lambda}_{\mathcal{N}} \end{pmatrix}\end{aligned}\tag{3}$$

In addition to the recursive expression for the orthogonal basis $\psi_k(x)$, a more elegant characterization for Legendre polynomials in terms of Bernstein basis functions is

$$\psi_k(x) = \sum_{j=0}^k (-1)^{k+j} \binom{k}{j} b_j^k(x).$$

Farouki (2000) studied transformations between the Legendre and Bernstein order- \mathcal{N} polynomials defined on $[0, 1]$, deriving explicit forms for the elements that relate $\eta_0^{\mathcal{N}}, \dots, \eta_{\mathcal{N}}^{\mathcal{N}}$ and $\lambda_0, \dots, \lambda_{\mathcal{N}}$ in (2) and (3) respectively, through the linear transformation $\eta_j^{\mathcal{N}} = \sum_{k=0}^{\mathcal{N}} Q_{jk}^{(\mathcal{N})} \lambda_k$. More generally $\begin{pmatrix} \eta_0 \\ \boldsymbol{\eta}_{\mathcal{N}} \end{pmatrix} = \mathbf{Q}_{\mathcal{N}} \begin{pmatrix} \lambda_0 \\ \boldsymbol{\lambda}_{\mathcal{N}} \end{pmatrix}$, where the coefficients $Q_{jk}^{(\mathcal{N})}$ of the transformation matrix $\mathbf{Q}_{\mathcal{N}}$

are defined as

$$Q_{jk}^{(\mathcal{N})} = \frac{1}{\binom{\mathcal{N}}{j}} \sum_{i=\max(0, j+k-\mathcal{N})}^{\min(j, k)} (-1)^{k+i} \left[\binom{k}{i} \right]^2 \binom{\mathcal{N}-k}{j-i}, \quad (4)$$

for $0 \leq j, k \leq \mathcal{N}$. Conversely, the Legendre coefficients may be derived from the Bernstein coefficients with the reverse mapping $\lambda_{\mathcal{N}} = \mathbf{Q}_{\mathcal{N}}^{-1} \eta_{\mathcal{N}}$, where the elements of $\mathbf{Q}_{\mathcal{N}}^{-1}$ are

$$\left(Q_{jk}^{(\mathcal{N})} \right)^{-1} = \frac{2j+1}{\mathcal{N}+j+1} \binom{\mathcal{N}}{k} \sum_{i=0}^j (-1)^{j+i} \frac{\binom{j}{i} \binom{j}{i}}{\binom{\mathcal{N}+j}{k+i}} \quad (5)$$

The expressions in (4) and (5) can also be computed recursively, avoiding the need to compute the combinatorial terms that may render the calculations numerically unstable.

Using the linear association between Bernstein and Legendre basis described in this section one may identify the order of smoothness for the Bernstein polynomial. In Section 3 we propose an alternative method, which consists of formulating a selection on a polynomial model space that obeys heredity constraints.

3 Methodology

The goal is to formulate a procedure with good operating properties to determine automatically a parsimonious (data-dependent) Bernstein polynomial approximation for $\mu(x)$. Suppose that for a random value of the predictor $X^* \sim X$ we want to predict a value from $Y^* = \mu(X^*) + \sigma\epsilon^*$, where $\mu(\cdot)$ is the true underlying value for the regression function and $\epsilon^* \sim \epsilon$. Now, for a given $k \in \{0, 1, \dots, \mathcal{N}\}$, note that Y^* under the order- k Bernstein polynomial model is given by $Y_k^* = \mu_k^{(\text{B})}(X^*) + \sigma\epsilon^*$, where $\mu_k^{(\text{B})}(\cdot)$ is defined as in (2), and denote by $\tilde{Y}_k^* = \mathbb{E}[Y_k^* | (x_1, y_1), \dots, (x_n, y_n)]$. The order of the optimal approximation $\mathcal{N}^* \in \{0, 1, \dots, \mathcal{N}\}$ (for $\mathcal{N} = \lfloor n^{2/3} \rfloor$) is defined as

$$\mathcal{N}^* = \underset{k \in \{0, \dots, \mathcal{N}\}}{\operatorname{argmin}} \mathbb{E} \left[\left(Y^* - \tilde{Y}_k^* \right)^2 \middle| (x_1, y_1), \dots, (x_n, y_n) \right], \quad (6)$$

where the posterior expectation is with respect to the posterior distribution of (X^*, Y^*) given the data $(x_1, y_1), \dots, (x_n, y_n)$. In other words, \mathcal{N}^* is the degree of the polynomial that minimizes the posterior predictive squared error loss.

To obtain \mathcal{N}^* , we formulate the nonparametric estimation problem with BPs in a way that makes tools from the objective Bayesian testing literature compatible with this problem. As discussed in Section 2, it is first necessary to move onto the Legendre basis, which is orthogonal, and preserves, as much as possible, the desirable features of the Bernstein form. In the Legendre space, once the parameter and model prior probabilities have been defined, the variable selection procedure can be implemented, provided a method for estimating the posterior probabilities is available. In the case of a polynomial basis generated from a single predictor, the size of the model space corresponds to the number of terms in the highest order model considered ($\mathcal{N} + 1$ for the prob-

lem at hand). Since the model space can be easily enumerated, model posterior probabilities can be calculated in closed form, precluding the need for a stochastic search algorithm to explore the model space. For problems where two or more predictors are considered, and a multivariate Bernstein polynomial is used for nonparametric estimation, the stochastic search algorithm proposed in Taylor-Rodriguez et al. (2015) may be used.

In the remainder of this section, we elaborate on how the Legendre representation enables choosing the order of smoothness via a Bayesian selection algorithm that respects the polynomial structure in the predictor space of the nonparametric models being considered.

To begin with, let model space \mathcal{M} (in its Legendre form) correspond to the set of models $\{\gamma_0, \gamma_1, \dots, \gamma_N\}$, where the order- k polynomial model is given by $\gamma_k = (\gamma_0, \gamma_{k,1}, \gamma_{k,2}, \dots, \gamma_{k,N})$, where $\gamma_{k,j} = 1$ if $j \leq k$ and $\gamma_{k,j} = 0$ for $j > k$. That is, the $\gamma_{k,j}$ is the inclusion indicator variable for the order- j term in γ_k . Hereinafter, we assume that all probabilistic statements are conditional on model space \mathcal{M} and for simplicity avoid conditioning on it explicitly.

In the Bayesian paradigm model $\gamma_k \in \mathcal{M}$ is characterized by a sampling density and a prior distribution. The sampling density associated with γ_k is $f(\mathbf{y}|\lambda_0, \boldsymbol{\lambda}_k, \tau, \gamma)$, where λ_0 , $\boldsymbol{\lambda}_k$, and τ are the parameter in the base model γ_0 , the vector of parameters included in γ_k different from λ_0 , and the precision parameter, respectively. The prior probability for model γ_k and its corresponding set of parameters is $\pi(\lambda_0, \boldsymbol{\lambda}_k, \tau, \gamma_k) = \pi(\lambda_0, \boldsymbol{\lambda}_k, \tau) \pi(\gamma_k)$. In this section we introduce the priors used on the parameters and the models.

3.1 Parameter priors: Mixtures of g -priors with normal response

Objective local priors for the model parameters $(\lambda_0, \boldsymbol{\lambda}_k, \tau)$ are achieved through modifications and extensions of Zellner's g -prior (Berger and Pericchi, 1996; Liang et al., 2008; Moreno et al., 1998; Womack et al., 2014). These are referred to as scaled mixtures of g -priors, and in general, they share similar good operating properties with slight differences in their behavior near the origin. The distribution assumed for the precision parameter (ω in the equations below) determines the type of mixture of g -prior being considered; the possible priors considered for this parameter are described in Section 3.3.

Let matrices $\boldsymbol{\Psi}_0$ and $\boldsymbol{\Psi}_k$ contain the Legendre bases (evaluated at the observed values for X) related to parameters λ_0 and $\boldsymbol{\lambda}_k$, respectively. The parameters λ_0 and τ that conform the base model γ_0 , are assigned a reference prior (e.g., Jeffreys prior) and are assumed to be within every model in \mathcal{M} . The likelihood function is given by $\mathbf{y}|\boldsymbol{\lambda}, \gamma_k \sim \mathcal{N}(\lambda_0 \boldsymbol{\Psi}_0 + \boldsymbol{\Psi}_k \boldsymbol{\lambda}_k, \tau \mathbf{I})$; letting $q_k = k + 1$ (i.e., the number of regression coefficients in γ_k), the form of the scaled mixtures of g -priors is

$$\pi(\lambda_0, \boldsymbol{\lambda}_k, \tau|\omega) = c_0 \tau \times \mathcal{N}_{q_k} \left(\boldsymbol{\lambda}_k \mid \mathbf{0}, \left(\frac{\omega \tau (q_k + 1)}{n} \boldsymbol{\Psi}_k' (\mathbf{I} - \mathbf{P}_0) \boldsymbol{\Psi}_k \right)^{-1} \right) \quad (7)$$

where $g = 1/\omega$ in the g -prior formulation, \mathbf{P}_0 is the hat matrix for $\boldsymbol{\Psi}_0$. Given the orthogonality of

the Legendre polynomial, $\pi(\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_k, \tau|\omega)$ further reduces to

$$\pi(\lambda_0, \boldsymbol{\lambda}_k, \tau|\omega) = c_0 \tau \times \mathcal{N}_{q_k} \left(\boldsymbol{\lambda}_k \mid \mathbf{0}, \left(\frac{\omega \tau (q_k + 1)}{n} \boldsymbol{\Psi}'_k \boldsymbol{\Psi}_k \right)^{-1} \right), \quad (8)$$

with $\boldsymbol{\Psi}'_k \boldsymbol{\Psi}_k = \text{diag} \left\{ \tilde{\boldsymbol{\psi}}'_j \tilde{\boldsymbol{\psi}}_j : \gamma_j \in \gamma \setminus \gamma_0 \right\}$, where $\tilde{\boldsymbol{\psi}}_j$ denotes the j th column of $\boldsymbol{\Psi}_k$, for n large and the predictor values in their original scale (i.e., x_1, \dots, x_n) are uniformly distributed. Although these assumptions might not strictly hold in practice, having a moderate to large sample size suffices for $\boldsymbol{\Psi}'_k \boldsymbol{\Psi}_k$ to be approximately diagonal.

3.2 Priors on the model space

Different approaches have been developed to account for structure in the predictor space within variable selection procedures (e.g., Bien et al., 2013; Chipman, 1996; Yuan et al., 2009). Among them, the methodology proposed by Chipman (1996) translates beliefs about model plausibility into prior distributions, thus, helping to account for the hierarchical structure among polynomial predictors.

The type of heredity constraints in model spaces with predictors that have a polynomial hierarchical structure can be translated into prior probability functions by relying on the inheritance principle. The inheritance principle assumes that the inclusion of a higher order term only depends on the inclusion indicators of lower-order terms from which it inherits (e.g., the inclusion of x_1^3 only depends on the inclusion of x_1 and x_1^2). The construction for the model prior probabilities can be further simplified by assuming *immediate inheritance*. This principle states that the inclusion of a given predictor (say x_1^3) conditional on the inclusion of its immediate ancestors (x_1^2 in this case), is independent from the inclusion of any other term in the model space.

As mentioned before, each model γ_k is associated to a binary vector $\boldsymbol{\gamma}_k = (1, \gamma_{k,1}, \dots, \gamma_{k,\mathcal{N}})$, with $\gamma_{k,j} = \mathbf{I}(j \leq k)$, representing the inclusion status for the $\mathcal{N} + 1$ polynomial terms, and the order 0 term, which is always included, in the order- k model. For the nonparametric problem in a single predictor, assuming conditional independence and immediate inheritance, and, following the approach of Taylor-Rodriguez et al. (2015), the model prior probabilities are

$$\pi(\boldsymbol{\gamma}_k) = \int \left(\prod_{j=1}^{\mathcal{N}} \pi_{k,j}^{\gamma_{k,j}} (1 - \pi_{k,j})^{1-\gamma_{k,j}} \right) p(\boldsymbol{\pi}) d\boldsymbol{\pi} \quad (9)$$

with $\gamma_{k,j}$ defined as above, $\pi_{k,j} = \Pr(\gamma_{k,j} = 1 | \gamma_{k,j-1})$, and $p(\boldsymbol{\pi}) = \prod_{k=1}^{\mathcal{N}} p_{\pi}(\pi_k)$ is the prior on the inclusion probabilities for each term, with

$$p_{\pi}(\pi_{k,j}) = \begin{cases} \text{Beta}(a_{k,j}, b_{k,j}) & \text{if } \gamma_{k,j-1} = 1 \\ 0 & \text{if } \gamma_{k,j-1} = 0 \end{cases} \text{ for } k = 1, \dots, \mathcal{N}.$$

The prior density $p_{\pi}(\pi_{k,j})$ reflects the fact that whenever the order $j - 1$ term is excluded, terms of orders j and higher should also be excluded. As such, the inclusion probabilities $\pi_{k,j}$ only

enable paths to models that respect the polynomial hierarchy.

3.3 Precision parameter and model posterior probabilities

The prior specification is completed by assigning a distribution to ω . Considering different priors on ω yields different objective priors. In particular, the intrinsic prior assumes that $\omega \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$. Also a version of the Zellner-Siow prior is given by $\omega \sim \text{Gamma}(\frac{\nu}{2}, \frac{\rho}{2})$ parameterized to have mean $\frac{\nu}{\rho}$, which produces a multivariate Cauchy distribution on $\boldsymbol{\lambda}$ when $\nu = \rho = 1$. And finally, a family of hyper- g priors is given by $\omega|\rho \sim \text{Gamma}(\frac{\nu}{2}, \frac{\rho}{2})$, where $\rho \sim \text{Gamma}(\frac{a}{2}, \frac{b}{2})$, with $\nu = 1, a = 2$, and $b = 1$ recommended. These latter ones have Cauchy-like tails but produce more shrinkage than the Cauchy prior (Womack et al., 2014).

Assuming the priors described above for the parameters and the models, the Bayes factor for γ_k relative to γ_0 is given by

$$BF_{\gamma_k, \gamma_0}(\mathbf{y}) = (1 - R_k^2)^{\frac{n-q_0}{2}} \int \left(\frac{n + \omega(q_k + 1)}{n + \frac{\omega(q_k+1)}{1-R_k^2}} \right)^{\frac{n-q_k}{2}} \left(\frac{\omega(q_k + 1)}{n + \frac{\omega(q_k+1)}{1-R_k^2}} \right)^{\frac{q_k-q_0}{2}} \pi(\omega) d\omega,$$

with $R_k^2 = \mathbf{y}'\Psi_k(\Psi_k'\Psi_k)^{-1}\Psi_k'\mathbf{y}/\mathbf{y}'(\mathbf{I} - \mathbf{P}_0)\mathbf{y}$.

The model posterior probabilities, conditional on the observed responses \mathbf{y} for a model $\gamma_k \in \mathcal{M}$ can be obtained in terms of its Bayes factor with respect to the base model γ_0 as

$$p(\gamma_k|\mathbf{y}) = \frac{BF_{\gamma_k, \gamma_0}(\mathbf{y})\pi(\gamma_k)}{\sum_{\gamma \in \mathcal{M}} BF_{\gamma, \gamma_0}(\mathbf{y})\pi(\gamma)}, \quad (10)$$

4 Optimality under the Median Probability Model

It is well known that the Bayesian model averaging estimators are optimal for prediction under squared error loss (Raftery et al., 1997). However, if a single model is of interest among a particular set of models, Barbieri and Berger (2004) demonstrated that in many scenarios the MPM is optimal for prediction under squared error loss. Their results rest on the assumption that the posterior mean for the parameters of any model inside model space \mathcal{M} , simply corresponds to the relevant coordinates from the posterior mean for the full model. Unfortunately, this assumption does not hold when using mixtures of g -priors, hence, the results obtained in Barbieri and Berger (2004) are not directly applicable. That being said, for the problem of interest, an analogous result can be ascertained asymptotically, as shown below.

First, let $\bar{\lambda}_0$ and $\bar{\lambda}$ represent the resulting model averaging parameter estimates for the parameters in the full model $\gamma_{\mathcal{N}}$. Now, denote the least squares estimate for $\boldsymbol{\lambda}_{\mathcal{N}}$ by $\hat{\boldsymbol{\lambda}}_{\mathcal{N}} = (\Psi_{\mathcal{N}}'\Psi_{\mathcal{N}})^{-1}\Psi_{\mathcal{N}}'\mathbf{y}$, and by $\hat{\boldsymbol{\lambda}}_k = (\Psi_k'\Psi_k)^{-1}\Psi_k'\mathbf{y}$, that for $\boldsymbol{\lambda}_k$. Additionally, let $\tilde{\boldsymbol{\lambda}}_k$ be the posterior expectation of $\boldsymbol{\lambda}_k$ under model γ_k . With mixtures of g -priors, this posterior expectation is given by $\tilde{\boldsymbol{\lambda}}_k = \xi_k \hat{\boldsymbol{\lambda}}_k$, where $\xi_k = \mathbb{E}_{\omega} \left[\frac{n}{n + \omega(q_k + 1)} | \mathbf{y}, \gamma_k \right]$ is the shrinkage for $\boldsymbol{\lambda}_k$ (Womack et al., 2014).

Given $x^* \in [0, 1]$, defining $Y^* = Y(x^*)$, and conditioning on model space \mathcal{M} , the model averaging prediction for a new response is given by $\bar{y}^* = \mathbb{E}_{Y^*|\mathbf{y}} [Y^*|\mathbf{y}]$, which depends on the \mathcal{N} -vector of Legendre basis functions of orders 1 and higher, denoted by $\boldsymbol{\psi}_{\mathcal{N}}^* = \boldsymbol{\psi}_{\mathcal{N}}(x^*)$. Finally, let \mathbf{H}_k be the $\mathcal{N} \times k$ matrix with elements $h_{jj} = 1$ and $h_{mj} = 0$, where $m \neq j$, $j = 1, \dots, k$ and $m = 1, \dots, \mathcal{N}$. Matrix \mathbf{H}_k identifies the active predictors in model γ_k , such that $\hat{\boldsymbol{\lambda}}_k = \mathbf{H}_k' \hat{\boldsymbol{\lambda}}_{\mathcal{N}}$ and $(\boldsymbol{\psi}_{\mathcal{N}}^*)' \mathbf{H}_k = (\boldsymbol{\psi}_k^*)'$. The model averaging predictor is then given by

$$\begin{aligned} \bar{y}^* &= \bar{\lambda}_0 + (\boldsymbol{\psi}_{\mathcal{N}}^*)' \bar{\boldsymbol{\lambda}} = \hat{\lambda}_0 + (\boldsymbol{\psi}_{\mathcal{N}}^*)' \left(\sum_{k=1}^{\mathcal{N}} p(\gamma_k|\mathbf{y}) \mathbf{H}_k \tilde{\boldsymbol{\lambda}}_k \right) \\ &= \hat{\lambda}_0 + (\boldsymbol{\psi}_{\mathcal{N}}^*)' \left(\sum_{k=1}^{\mathcal{N}} p(\gamma_k|\mathbf{y}) \mathbf{H}_k \left(\xi_k \hat{\boldsymbol{\lambda}}_k \right) \right) \\ &= \hat{\lambda}_0 + (\boldsymbol{\psi}_{\mathcal{N}}^*)' \left(\sum_{k=1}^{\mathcal{N}} \xi_k p(\gamma_k|\mathbf{y}) \mathbf{H}_k \mathbf{H}_k' \right) \hat{\boldsymbol{\lambda}}_{\mathcal{N}}, \end{aligned} \quad (11)$$

where $p(\gamma_k|\mathbf{y})$ is the posterior probability of some model $\gamma_k \in \mathcal{M}$.

Now, define $p_j = \sum_{k=1}^{\mathcal{N}} p(\gamma_k|\mathbf{y}) \gamma_{k,j}$ and $\tilde{p}_j = \sum_{k=1}^{\mathcal{N}} \xi_k p(\gamma_k|\mathbf{y}) \gamma_{k,j}$ with $j = 1, \dots, \mathcal{N}$. Recall that \mathcal{N} is the number of terms in $\gamma_{\mathcal{N}}$ excluding the order 0 term. Lastly, let $\tilde{y}_k^* = \tilde{\mu}_k(x^*)$, which is the posterior predictive mean conditional on $X^* = x^*$ under model γ_k . Hence, given the model space \mathcal{M} , the posterior predictive loss with respect to model γ_k for a new response $Y^* = Y(x^*)$, is given by

$$\begin{aligned} \mathbf{L}_{\mathcal{N}}(\gamma_k, x^*) &= \mathbb{E}_{Y^*|\mathbf{y}} [(Y^* - \tilde{y}_k^*)^2 | \mathbf{y}] = \mathbb{E}_{Y^*|\mathbf{y}} [(Y^* - \bar{y}^* + \bar{y}^* - \tilde{y}_k^*)^2 | \mathbf{y}] \\ &= \text{Var}(Y^*|\mathbf{y}) + (\bar{y}^* - \tilde{y}_k^*)^2 \\ &= \text{Var}(Y^*|\mathbf{y}) + \sum_{j=1}^{\mathcal{N}} \left(\hat{\lambda}_j d_j \right)^2 (\tilde{p}_j - \xi_k \gamma_{k,j})^2, \end{aligned} \quad (12)$$

where d_j represents the j th diagonal element of $\boldsymbol{\psi}^* (\boldsymbol{\psi}^*)'$, which for large n can be approximately equal to $\tilde{\boldsymbol{\psi}}_j' \tilde{\boldsymbol{\psi}}_j$, the j th diagonal element of $\boldsymbol{\Psi}' \boldsymbol{\Psi}$. The optimal model is the one that minimizes $\mathbb{E} [\mathbf{L}_{\mathcal{N}}(\gamma_k, X^*)]$, where the expectation is over $X^* \sim X$. Note, that the choice of \mathcal{N}^* only depends on \mathbf{y} , and that this minimization amounts to minimizing the second term on the right hand side of (12), as $\text{Var}(Y^*|\mathbf{y})$ is not a function of γ_k . Below we demonstrate that when using an orthogonal basis in conjunction with mixtures of g-priors for the nonparametric regression problem, this minimization is asymptotically attained by the MPM.

Theorem 1 *Assuming a mixture of g-priors for the model parameters, the order chosen by the median probability model is asymptotically optimal for prediction, under the loss function in (12), if $\frac{\mathcal{N}}{n} \rightarrow 0$ as $n \rightarrow \infty$ and $\mathbb{E} [\omega|\mathbf{y}, \gamma_{\mathcal{N}}] \xrightarrow{a.s.} \omega_0 \in [0, \infty)$.*

Proof For every model in \mathcal{M} and any observed response vector \mathbf{y} , the shrinkage ξ_k is bounded above by 1. Additionally, since the shrinkage is decreasing in the number of predictors then $\xi_{\mathcal{N}} \leq$

$\dots \leq \xi_1 \leq \xi_0 \leq 1$. By Jensen's inequality, we have that the shrinkage for the full model is such that

$$\xi_{\mathcal{N}} = \mathbb{E} \left[\frac{n}{n + \omega(\mathcal{N} + 2)} \middle| \mathbf{y}, \gamma_{\mathcal{N}} \right] \geq \frac{1}{1 + \frac{\mathcal{N}+2}{n} \mathbb{E} [\omega | \mathbf{y}, \gamma_{\mathcal{N}}]}$$

where the expression on the right hand side of the inequality tends to 1 a.s. as $n \rightarrow \infty$ if $\mathcal{N}/n \rightarrow 0$ and $\mathbb{E} [\omega | \mathbf{y}, \gamma_{\mathcal{N}}] \rightarrow \omega_0 \in [0, \infty)$. Hence, $\xi_k \xrightarrow{a.s.} 1$ for all $k = 0, 1, \dots, \mathcal{N}$.

Let $L_{\mathcal{N}}^0(\gamma_k, x^*) = \text{Var}(Y^* | \mathbf{y}) + \sum_{j=1}^{\mathcal{N}} (\hat{\lambda}_j d_j)^2 (p_j - \gamma_{k,j})^2$, which is minimized at the MPM for all n . Now, consider the difference

$$L_{\mathcal{N}}(\gamma_k, x^*) - L_{\mathcal{N}}^0(\gamma_k, x^*) = \sum_{j=1}^{\mathcal{N}} (\hat{\lambda}_j d_j)^2 [(\tilde{p}_j - \xi_k \gamma_{k,j})^2 - (p_j - \gamma_{k,j})^2],$$

which tends to 0 a.s. for all x^* as $n \rightarrow \infty$, since $\xi_k \xrightarrow{a.s.} 1$ for all $\gamma \in \mathcal{M}$. Hence, the loss function in (12) is asymptotically equivalent to $L_{\mathcal{N}}^0(\gamma_k, x^*)$ for each x^* . Next, by the Dominated Convergence Theorem, it follows that $\mathbb{E} [L_{\mathcal{N}}(\gamma_k, X^*)] \approx \mathbb{E} [L_{\mathcal{N}}^0(\gamma_k, X^*)]$. Finally, by the extension of the Argmax-Continuous Mapping Theorem derived by Ferger (2004), it follows that the minimizer of $\mathbb{E} [L_{\mathcal{N}}(\gamma_k, X^*)]$ will also converge a.s. to the minimizer of $\mathbb{E} [L_{\mathcal{N}}^0(\gamma_k, X^*)]$, hence the order chosen by the MPM is optimal for prediction (QED).

The conditions under which Theorem 1 holds are quite general, as $\mathbb{E} [\omega | \mathbf{y}, \gamma_{\mathcal{N}}]$ commonly converges to a finite non-negative value by the usual consistency results (e.g., see Liang et al., 2008; Womack et al., 2014), and additionally, from Tenbusch (1997) we know that \mathcal{N} should be at most $O(n^{2/3})$, such that the condition $\mathcal{N}/n \rightarrow 0$ is met in practice. Although \mathcal{N}^* as defined in (6) may be difficult to compute, by using Theorem 1, we obtain an approximate order $\hat{\mathcal{N}}^*$ by using the MPM. Next we explore the performance of $\hat{\mathcal{N}}^*$ using several simulated data scenarios.

5 Simulation Study

In this section we test the ability of the proposed selection algorithm, choosing the order of the BPs using several simulated data scenarios. Additionally, we contrast these results to those obtained by 5-fold cross validation (CV) with functions from the `cvTools` package in R.

In particular, we ran two distinct sets of simulations. The first one takes into account all combinations of sample size ($n = 100, 200, 500$), signal-to-noise ratio (SNR=0.5, 1, 2 defined later) and two true mean functions $\mu(\cdot)$. With each combination of these, we drew at random 100 datasets with one predictor and its corresponding response. The second set of simulations is aimed at determining whether or not, when compared to cross-validation, the proposed strategy yields similar models in terms of accuracy and parsimony, and if our method is computationally as efficient. For this set of simulations 100 datasets were drawn assuming $n = 10^4$, SNR= 2 and only one true mean function.

For each simulated dataset, we first generated $x_i \sim \text{U}[a, b]$, and then the response $y_i \sim \text{N}(\mu(x_i), \sigma^2)$, with σ determined by the SNR level chosen. The definition considered for the SNR, with $x \in [a, b]$, is

$$\text{SNR} = \frac{1}{\sigma} \int_a^b \frac{|\mu(x)|}{b-a} dx.$$

The first mean function relates the outcome to the predictor through an order five polynomial. The second, corresponds to a piecewise linear function, which is more challenging as it is not differentiable everywhere. Both our method and cross-validation were applied to every dataset to select the order of smoothness. In every case we assumed the full model (largest possible model) to be of order $\mathcal{N} = \lfloor n^{2/3} \rfloor$. The results were assessed in terms of:

Computational speed: the time (in seconds) taken by each method per dataset

Model complexity: frequency counts for the selected order (out of the 100 datasets per scenario) with each method

Accuracy: the sup-norm, given by $\kappa_k = \|\tilde{\mu}_k(\cdot) - \mu(\cdot)\|_\infty$, as defined in Section 2. This norm was chosen as it dominates all other L_p norms (i.e., $\|\tilde{\mu}(\cdot) - \mu(\cdot)\|_p \leq \|\tilde{\mu}(\cdot) - \mu(\cdot)\|_\infty$ for $p \geq 1$). The sup-norms considered were

- $\|\tilde{\mu}(\cdot) - \tilde{\mu}_{\mathcal{N}}(\cdot)\|_\infty$, to see the performance of the largest order model considered;
- $\|\tilde{\mu}_{\mathcal{N}^*}(\cdot) - \mu(\cdot)\|_\infty$, where \mathcal{N}^* corresponds to the order of the MPM resulting from the selection;
- $\|\tilde{\mu}_{\mathcal{N}^{\text{cv}}}(\cdot) - \mu(\cdot)\|_\infty$, where \mathcal{N}^{cv} is the order of smoothness chosen by 5-fold cross-validation.

5.1 Overall performance

The first mean function used is the order 5 polynomial given by $\mu(x) = 5x(5x - 0.2)(0.4x - 1.8)(3x - 1.8)(2x - 1.8)$, and the second one is the piece-wise linear function $\mu(x) = x\text{I}_{[-3,-1]}(x) - \text{I}_{[-1,1]}(x) + (x-2)\text{I}_{[1,3]}(x)$ (see Figure 1). With both functions, the fitted curves for the 100 datasets generated from these models matched closely the true curves, regardless of the sample sizes and levels of SNR (see Figures 9 and 10 in the appendix). For the particular order 5 polynomial chosen, the true mean function has three stationary points, implying that it can be approximated by an order four polynomial. Interestingly, in this case the MPM commonly chose the order to be 4 for all sample sizes and SNR values (see Figure 5). In spite of the fact that this is not the true model, this indicates that the method will identify a suitably parsimonious approximation. Similar values for the order of smoothness were often selected with CV; however, as the SNR or sample size increased, for some datasets CV chose polynomials of substantially larger degree. For the piece-wise linear function, as the sample size or the SNR increased with either method the selected order increased, again spreading out to slightly larger values when using CV (Figure 7 in appendix).

Even though CV is selecting more complex models as SNR or n grow with both mean functions $\mu(\cdot)$ used in our simulations, the accuracy of these models does not improve that from the MPM's, as evidenced in Figures 6 and 8 in the Appendix. Finally, in either case, it is also clear that simply using the full order \mathcal{N} model overfits the data, as it produces the largest predictive errors among the three orders of smoothness compared, namely \mathcal{N}^* , \mathcal{N}^{cv} and \mathcal{N} . Lastly, the computational gains from using the proposed approach when compared to CV are astonishing, reducing the computational time by one or two orders of magnitude (see Tables 1 and 2 in the Appendix).

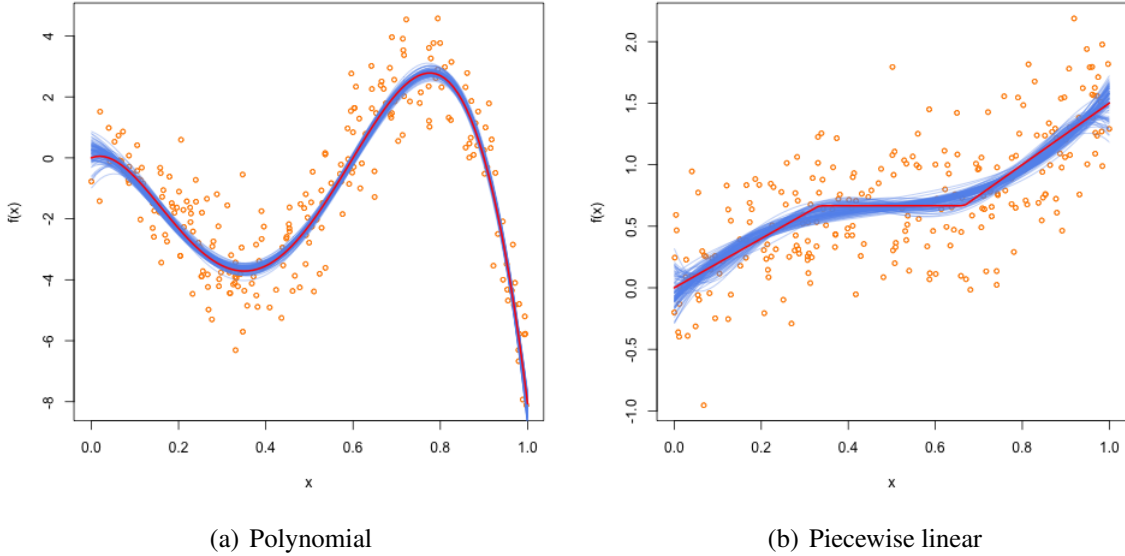
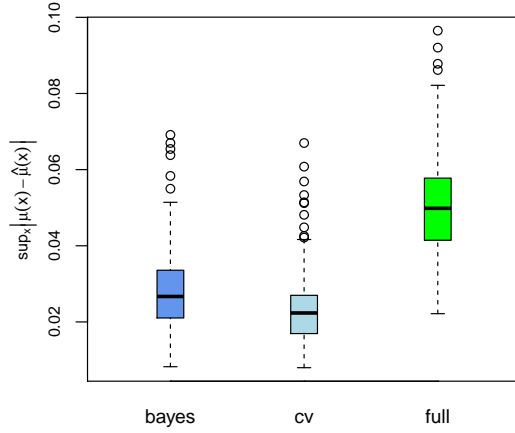


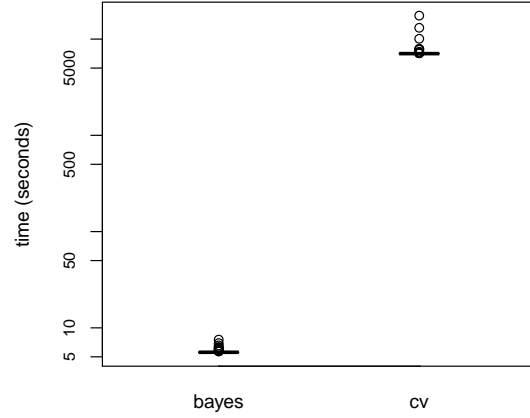
Figure 1: Functions considered for simulated data experiments. True mean function in red, simulated data in orange, and fitted curves using MPM in blue

5.2 Comparison with cross-validation under ideal conditions

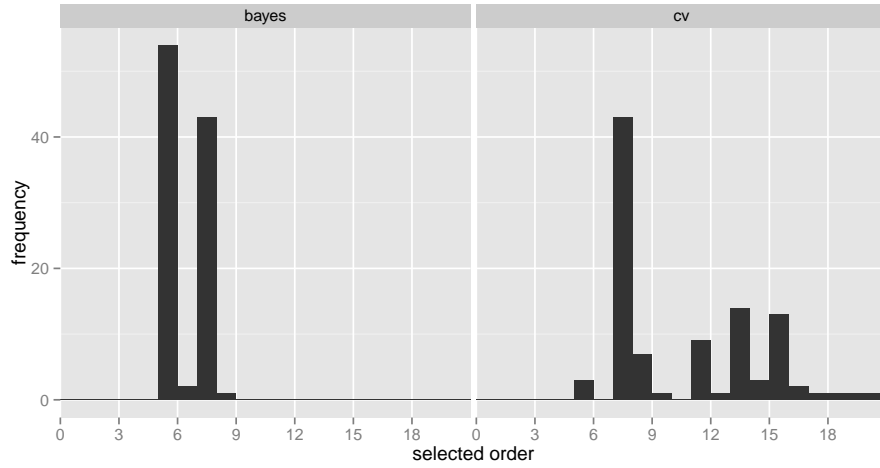
Given the results from the set of simulations described in the previous section, the intent behind this set of simulation is to settle if there are in fact any significant differences between the results obtained choosing the order using the MPM and that chosen by CV. The conditions used for this simulation were ideal in terms of having a very large sample size and considerably large signal so that good models would be easily identified. Again, we used the piece-wise linear function $\mu(x) = xI_{[-3,-1]}(x) - I_{[-1,1]}(x) + (x-2)I_{[1,3]}(x)$. The conclusions are unequivocal: the proposed method, when compared to CV, is comparable in predictive ability, chooses more parsimonious models, and does so in a fraction of the time taken by CV (Figure 2). In addition, the results for the full order \mathcal{N} were also included in our analysis providing clear indication of data overfitting, this was a recurring theme in all of the simulation experiments performed throughout the section.



(a) sup-norm



(b) computation time



(c) selected order of smoothness

Figure 2: sup-norm error, computation time and frequencies for the selected order of the mean function $\mu(x) = xI_{[-3,-1]}(x) - I_{[-1,1]}(x) + (x - 2)I_{[1,3]}(x)$ in 100 datasets, using the proposed method (bayes) and cross-validation (cv).

6 Data examples

In this section, we select the order of smoothness \mathcal{N}^* for the estimated mean function $\tilde{\mu}_{\mathcal{N}^*}(\cdot)$ with two real datasets using Bernstein polynomials. One of the datasets has continuous response and the other one has a binary one. We also describe succinctly the strategy used to adapt the method for binary responses. It is important to highlight that in this section we provide the distribution for the order of smoothness, with which we can quantify the uncertainty associated to selected order. This is not possible using CV; in the previous section we were able to provide frequency plots given that each scenario was analyzed on 100 datasets.

6.1 Selecting \mathcal{N}^* with continuous response

For this example we use the child growth data from Ramsay (1998), which was also previously analyzed in Curtis and Ghosh (2011). These data contain 83 height measurements from a 10 year old boy over a period of 312 days. In this analysis, we compare the selected order \mathcal{N}^* BP model, to the mean function estimated using Bayesian isotonic regression method (Curtis and Ghosh, 2011) found in the `bisoreg` package in R.

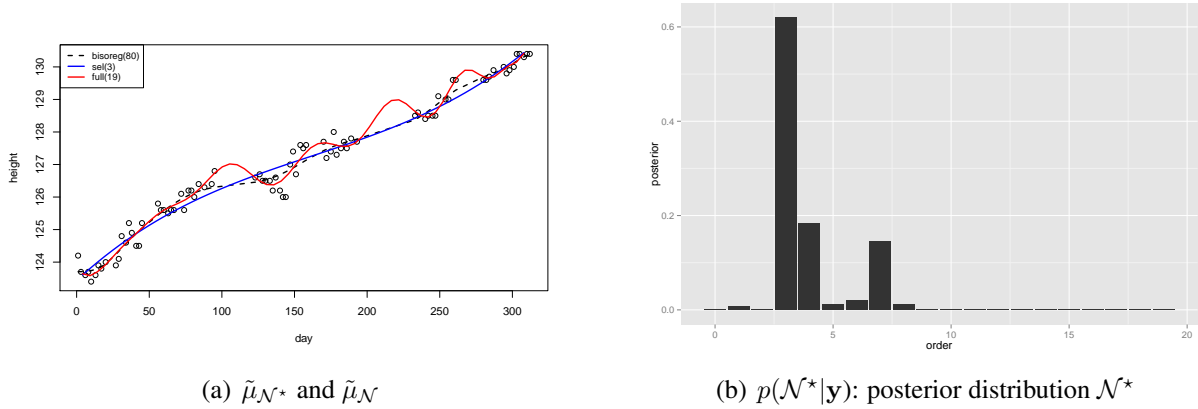


Figure 3: Child height vs days. Lines in blue and red correspond to the fitted means using the order \mathcal{N}^* and \mathcal{N} BP models and the dashed line corresponds to fit using Bayesian isotonic regression with BPs (left). Posterior probability for the order of smoothness \mathcal{N}^* (right).

In this example there is a fair degree of certainty in the choice for the order of smoothness, as more than 60% of the distribution concentrates on $\mathcal{N}^* = 3$ when we choose $\mathcal{N} = \lfloor n^{2/3} \rfloor = 20$. A reasonable assumption for this type of data is that it is non-decreasing; however, the full model fit again goes against common sense. As such, we compare the results from our procedure to those from an order-80 Bayesian monotonic BP regression. Remarkably, the parsimonious model selected with our method, provides a fit very close to that of `bisoreg`, with $\sup_{x \in [0,1]} |\tilde{\mu}_{\mathcal{N}^*}(x) - \tilde{\mu}_{\text{bisoreg}}(x)| = 0.2541$, which is outstanding considering that the observed range for the responses is between 123 and 131.

6.2 Selecting \mathcal{N}^* with binary response

In order to deal with binary responses we follow the strategy proposed in Leon-Novelo et al. (2012) with intrinsic priors, which uses data augmentation to introduce latent normal random variables that make the problem tractable. The approach builds upon the sampling algorithm of Albert and Chib (1993) for probit regression models. This selection problem is solved on the latent scale, since each model $\gamma \in \mathcal{M}$ can be viewed as a normal regression model where only the sign of the response is observed. Explicitly, associated to the response vector $\mathbf{y} = (y_1, \dots, y_n)$, where $y_i | \gamma_k \sim \text{Bernoulli}(\Phi(\boldsymbol{\psi}_k(x_i)' \boldsymbol{\lambda}_k))$, we may consider a sample of latent random variables $\mathbf{v} = (v_1, \dots, v_n)$, where $v_i | \gamma_k \sim N(v_i | \boldsymbol{\psi}_k(x_i)' \boldsymbol{\lambda}_k, 1)$. Each of the observed responses and the latent variables are connected through the equality $y_i = I(v_i > 0)$.

In the binary response case the goal is to derive the marginal probabilities for the observed responses from the marginal densities calculated for the latent variables. The marginal density of \mathbf{y} under model γ_k , are given by

$$\begin{aligned} m(\mathbf{y} | \gamma_k) &= \int_{A_1 \times \dots \times A_n} m(\mathbf{v} | \gamma_k) d\mathbf{v} \\ &= \int_{A_1 \times \dots \times A_n} \left(\int_{-\infty}^{\infty} m(\mathbf{v} | \gamma_k, \lambda_0) d\lambda_0 \right) d\mathbf{v} \\ &= \int_{-\infty}^{\infty} \underbrace{\left(\int_{\mathbf{A}} m(\mathbf{v} | \gamma_k, \lambda_0) d\mathbf{y} \right)}_{g_k(\mathbf{A} | \gamma_k, \lambda_0)} d\lambda_0 \end{aligned} \quad (13)$$

where $\mathbf{A} = A_1 \times \dots \times A_n$ and $A_i = (0, \infty)$ if $y_i = 1$ and $A_i = (-\infty, 0]$ otherwise. From equation 13 it follows that the Bayes factor for each model $\gamma \in \mathcal{M}$ in the binary case is given by

$$BF_{\gamma_k, \gamma_0}(\mathbf{y}) = \frac{\int_{-\infty}^{\infty} g_k(\mathbf{A} | \gamma_k, \lambda_0) d\lambda_0}{\int_{-\infty}^{\infty} g_0(\mathbf{A} | \gamma_0, \lambda_0) d\lambda_0}, \quad (14)$$

with

$$\begin{aligned} g_k(\mathbf{A} | \gamma_k, \lambda_0) &= c_0 \int_{\mathbf{A}} \mathcal{N}_n(\mathbf{v} | \lambda_0 \mathbf{1}_n, \Sigma_k) d\mathbf{y}, \\ g_0(\mathbf{A} | \gamma_0, \lambda_0) &= c_0 \int_{\mathbf{A}} \mathcal{N}_n(\mathbf{v} | \lambda_0 \mathbf{1}_n, \mathbf{I}_n) d\mathbf{y}, \end{aligned}$$

where $\Sigma_k = \mathbf{I}_n + \frac{2n}{k+1} \boldsymbol{\Psi}_k (\boldsymbol{\Psi}_k' \boldsymbol{\Psi}_k)^{-1} \boldsymbol{\Psi}_k'$. The model posterior probabilities are then obtained plugging in the Bayes factor as in Equation (10).

To illustrate the effectiveness of the method using the strategy of Leon-Novelo et al. (2012), we consider data from the experiment described in Martinez et al. (2012). This dataset relates the incidence of metritis in lactating cows, an early postpartum disease, to the levels of calcium in the blood. In this experiment, the authors were interested in demonstrating that subclinical hypocalcemia in dairy cows (defined as $\text{Ca} < 8.59 \text{ mg/dL}$) increases the likelihood of disease, and

in particular of metritis, during the early postpartum. The dataset includes measurements taken the third day after parturition from 73 multiparous cows in a dairy farm.

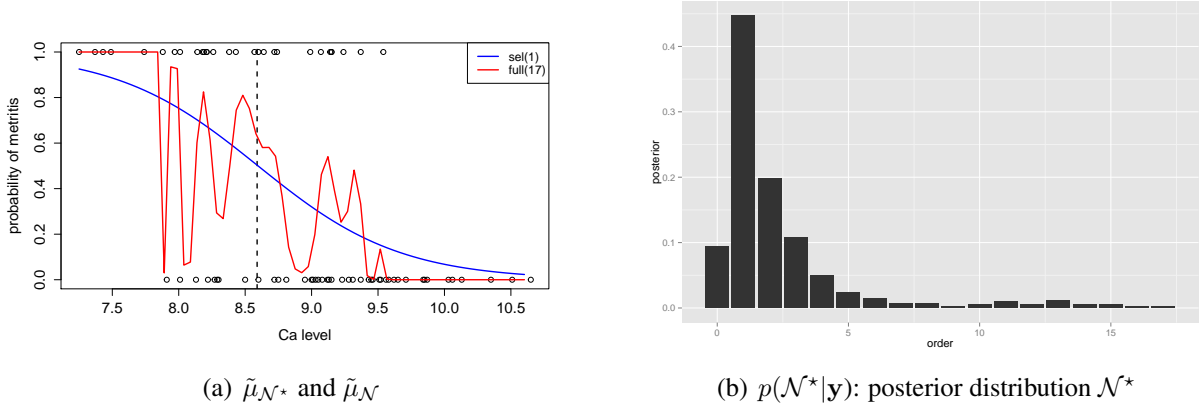


Figure 4: Probability of metritis vs level of Calcium in blood: Fitted means using the selected order \mathcal{N}^* model and the full order \mathcal{N} model (left). Posterior probability for the order of smoothness \mathcal{N}^* (right).

Figure 4 clearly demonstrates the inadequacy of using the full order \mathcal{N} Bernstein polynomial, which overfits the data, whereas the order \mathcal{N}^* BP model (in this case order 1 in probit scale) is consistent with the intuition that the relationship between calcium and metritis is negative monotonic. Interestingly, the point of inflection for the fitted curve with the selected order of smoothness occurs near the cutoff for subclinical hypocalcemia (dashed line in Figure 4). In spite of these results, the heavy tails of the posterior distribution for \mathcal{N}^* indicate that there is some uncertainty surrounding this choice.

7 Concluding Remarks

Using tools from the objective Bayesian variable selection literature and exploiting the fact that there is a linear map between Bernstein and orthogonal polynomial basis, we devise an efficient procedure to select optimally and automatically the order of smoothness used in the estimation of a nonparametric regression function with Bernstein Polynomials. The functions used throughout were built into the R package **AutoBPFit**. The proposed strategy was shown to be asymptotically optimal for prediction if the order of smoothness is chosen based on the MPM. In simulation experiments, our method, when compared to CV proved to be one or two orders of magnitude faster and selected consistently more parsimonious models, remarkably, without any significant loss in predictive accuracy. The method was extended to binary responses, and it can potentially be used with other exponential family responses where a latent normal representation is available. Throughout this article we treated the single predictor case, making use of the immediate inheritance assumption in the construction of the model priors. However, the method can be seamlessly extended to the multiple predictor case by enforcing the principle of *conditional independence* as

defined in Chipman (1996). In the case where several predictors are considered and selection must not only be done on the order of smoothness, but also on which predictors actually enter the model, it is commonly not possible to enumerate the entire model space, and instead a stochastic search algorithm to explore it can be considered (e.g., see Taylor-Rodriguez et al., 2015).

References

- Albert, J. H. and Chib, S. (1993). Bayesian-analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.
- Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, 32(3):870–897.
- Berens, H. and Lorentz, G. (1972). Inverse theorems for Bernstein polynomials. *Indiana University*.
- Berger, J. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.
- Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-section inference. *Annals of Statistics*, 41(2):802–837.
- Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141.
- Bouezmarni, T. and Rolin, J. M. (2007). Bernstein estimator for unbounded density function. *Journal of Nonparametric Statistics*, 19(3):145–161.
- Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36.
- Craven, P. and Wahba, G. (1978). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik*.
- Curtis, S. and Ghosh, S. K. (2011). A variable selection approach to monotonic regression with Bernstein polynomials. *Journal of Applied Statistics*, 38(5):961–976.
- Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121.
- Eilers, P. H. C. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:637–653.
- Eubank, R. (1985). Diagnostics for Smoothing Splines. *Journal of the Royal Statistical Society. Series B*, 47:332–341.
- Farouki, R. T. (2000). Legendre Bernstein basis transformations. *Journal of Computational and Applied Mathematics*, 119:145–160.
- Farouki, R. T. (2012). The Bernstein polynomial basis: A centennial retrospective. *Computer Aided Geometric Design*, 29(6):379–419.

- Ferger, D. (2004). A continuous mapping theorem for the argmax-functional in the non-unique case. *Statistica Neerlandica*, 58(1):83–96.
- Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19:1–67.
- Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Verlag, Berlin.
- Kelisky, R. P. and Rivlin, T. J. (1967). Iterates of Bernstein polynomials. *Pacific J. Math*, 21(3):511–520.
- Leblanc, A. (2010). A bias-reduced approach to density estimation using Bernstein polynomials. *Journal of Nonparametric Statistics*, 22(4):459–475.
- Lee, B. G., Park, Y., and Yoo, J. (2002). Application of Legendre-Bernstein basis transformations to degree elevation and degree reduction. *Computer Aided Geometric Design*, 19:709–718.
- Leon-Novelo, L., Moreno, E., and Casella, G. (2012). Objective Bayes model selection in probit models. *Statistics in medicine*, 31(4):353–65.
- Liang, F., Paulo, R., Molina, G., Clyde, M. a., and Berger, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, 103(481):410–423.
- Manté, C. (2015). Iterated Bernstein operators for distribution function and density estimation: Balancing between the number of iterations and the polynomial degree. *Computational Statistics & Data Analysis*, 84:68–84.
- Martinez, N., Risco, C. a., Lima, F. S., Bisinotto, R. S., Greco, L. F., Ribeiro, E. S., Maunsell, F., Galvão, K., and Santos, J. E. P. (2012). Evaluation of periparturient calcium status, energetic profile, and neutrophil function in dairy cows at low or high risk of developing uterine disease. *Journal of Dairy Science*, 95(12):7158–7172.
- Moreno, E., Bertolino, F., and Racugno, W. (1998). An intrinsic limiting procedure for model selection and hypotheses testing. *Journal of the American Statistical Association*, 93(444):1451–1460.
- Nadaraya, E. A. (1964). On Estimating Regression. *Theory of Probability & Its Applications*, 9:141–142.
- Osman, M. and Ghosh, S. K. (2012). Nonparametric regression models for right-censored data using Bernstein polynomials. *Computational Statistics and Data Analysis*, 56(3):559–573.
- Raftery, A., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56:611–630.
- Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):365–375.
- Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric regression*. Number 12. Cambridge university press.
- Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47:1–52.
- Stadtmüller, U. (1986). Asymptotic properties of nonparametric curve estimations. *Periodica Mathematica Hungarica*, 17(2).
- Taylor-Rodriguez, D., Womack, A., and Bliznyuk, N. (2015). Bayesian Variable Selection on Model Spaces Constrained by Heredity Conditions. *forthcoming in the Journal of Computational and Graphical Statistics* (<http://arxiv.org/abs/1312.6611>).
- Tenbusch, A. (1997). Nonparametric Curve Estimation With Bernstein Estimates. *Metrika*, pages 1–30.
- Wang, J. and Ghosh, S. (2012). Shape restricted nonparametric regression with Bernstein polynomials. *Computational Statistics & Data Analysis*, 56(9):2729–2741.
- Wang, Y. (2002). Interpolating Cubic Splines. *Journal of the American Statistical Association*, 97:366–366.
- Womack, A. J., León-Novelo, L., and Casella, G. (2014). Inference from Intrinsic Bayes Procedures Under Model Selection and Uncertainty. *Journal of the American Statistical Association*, (June):140114063448000.
- Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65:95–114.
- Yuan, M., Joseph, V. R., and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4):1738–1757.

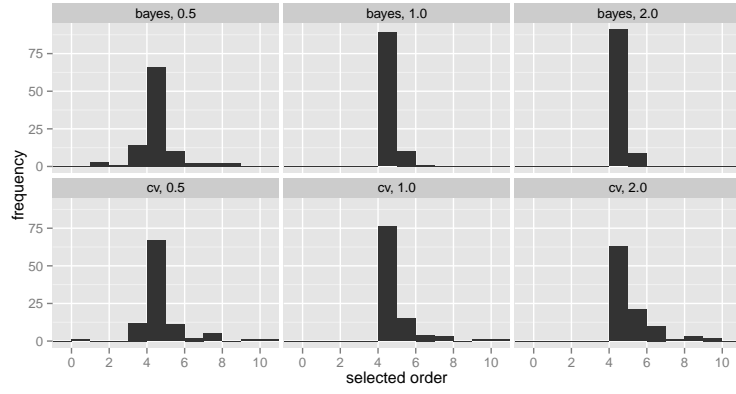
A Figures and Tables simulations Section 4.1

method	snr	$n = 100$			$n = 200$			$n = 500$		
		2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
bayes	0.5	0.07	0.08	0.09	0.08	0.10	0.11	0.10	0.12	0.19
cv		2.03	2.11	2.26	4.48	5.17	6.84	7.12	8.17	13.40
bayes	1	0.11	0.12	0.15	0.11	0.13	0.16	0.12	0.14	0.17
cv		2.01	2.18	2.97	3.32	3.63	5.12	7.39	9.09	20.77
bayes	2	0.08	0.12	0.15	0.08	0.09	0.11	0.06	0.08	0.09
cv		1.77	2.03	3.07	4.52	5.18	6.66	5.38	5.95	8.03

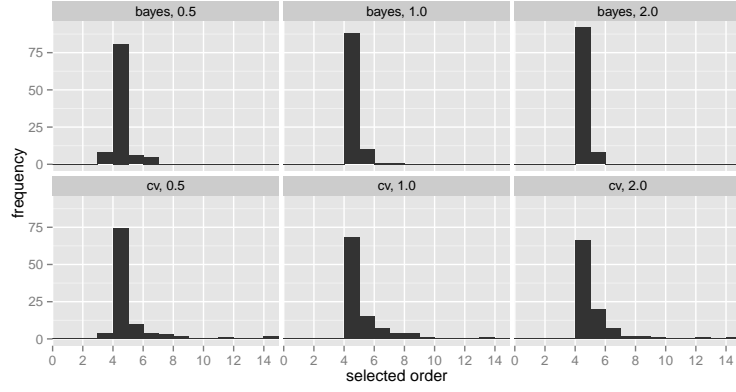
Table 1: Quantiles 2.5%, 50%, 97.5% for the computation time per dataset (in seconds). Comparison between BP order selection with proposed method (bayes) and cross-validation (cv) with mean function $\mu(x) = 5x(5x - 0.2)(0.4x - 1.8)(3x - 1.8)(2x - 1.8)$.

method	snr	$n = 100$			$n = 200$			$n = 500$		
		2.5%	50%	97.5%	2.5%	50%	97.5%	2.5%	50%	97.5%
bayes	0.5	0.07	0.08	0.09	0.12	0.14	0.16	0.12	0.14	0.17
cv		2.03	2.10	2.28	4.05	4.99	6.87	7.79	8.99	20.83
bayes	1	0.11	0.12	0.15	0.08	0.10	0.11	0.06	0.08	0.09
cv		2.03	2.17	2.98	4.48	5.19	6.76	5.40	5.97	7.90
bayes	2	0.06	0.08	0.09	0.11	0.12	0.16	0.07	0.07	0.08
cv		2.03	2.12	2.25	3.33	3.67	5.24	5.21	5.38	6.64

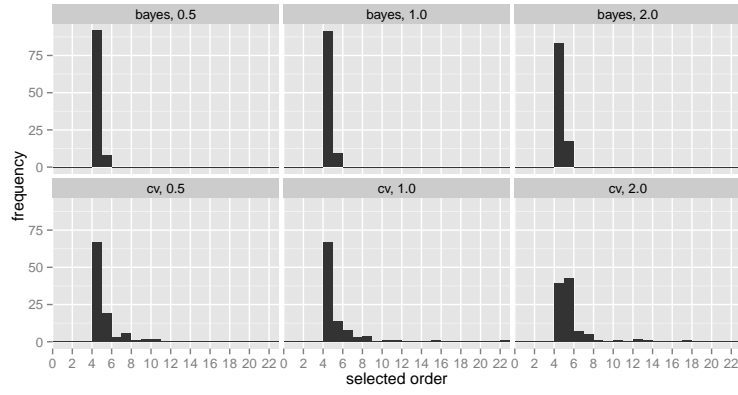
Table 2: Quantiles 2.5%, 50%, 97.5% for the computation time per dataset (in seconds). Comparison between BP order selection with proposed method (bayes) and cross-validation (cv) with mean function $\mu(x) = x\mathbf{I}_{\{x \in [-3, -1]\}} - \mathbf{I}_{\{x \in \{-1, 1\}\}} + (x - 2)\mathbf{I}_{\{x \in [1, 3]\}}$.



(a) $n = 100$

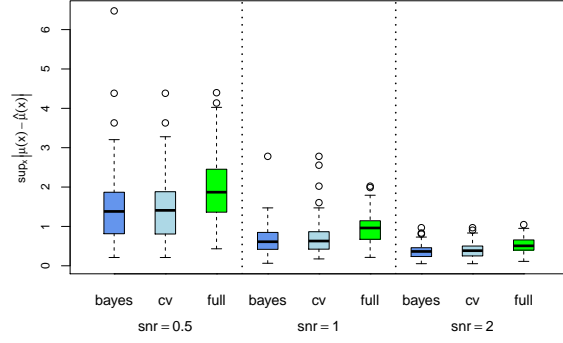


(b) $n = 200$

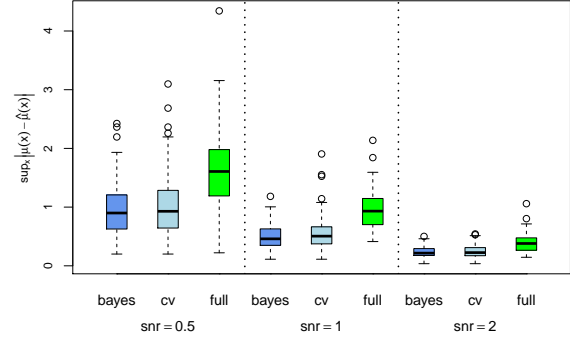


(c) $n = 500$

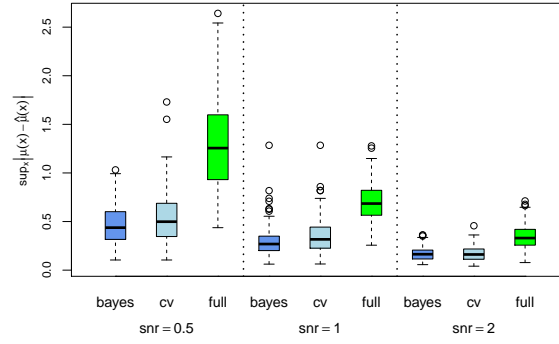
Figure 5: Frequencies for the order of smoothness selected in 100 simulated datasets with $SNR = 0.5, 1, 2$ for the proposed method (bayes) and cross-validation (cv) with mean function $\mu(x) = 5x(5x - 0.2)(0.4x - 1.8)(3x - 1.8)(2x - 1.8)$.



(a) $n = 100$

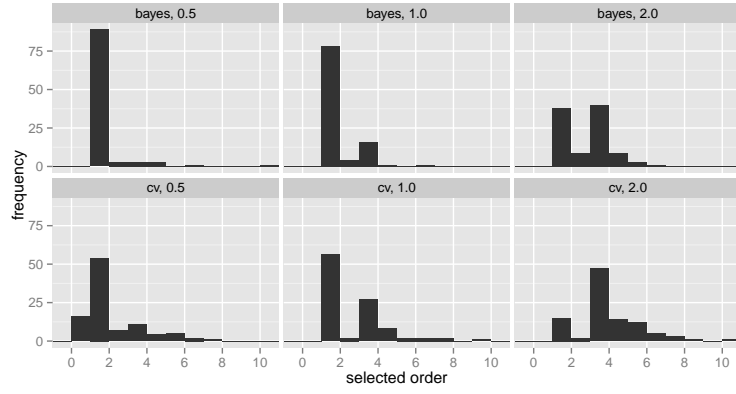


(b) $n = 200$

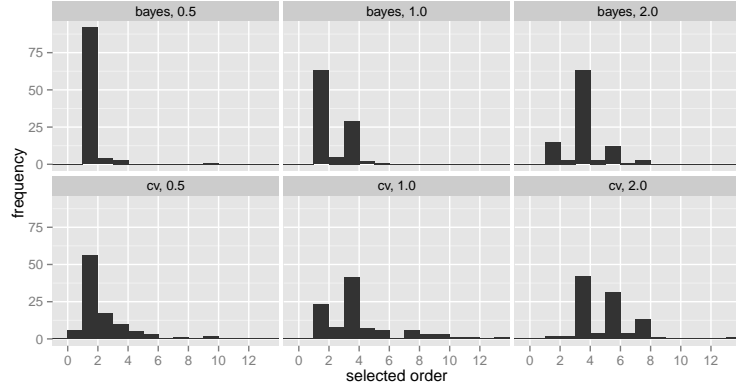


(c) $n = 500$

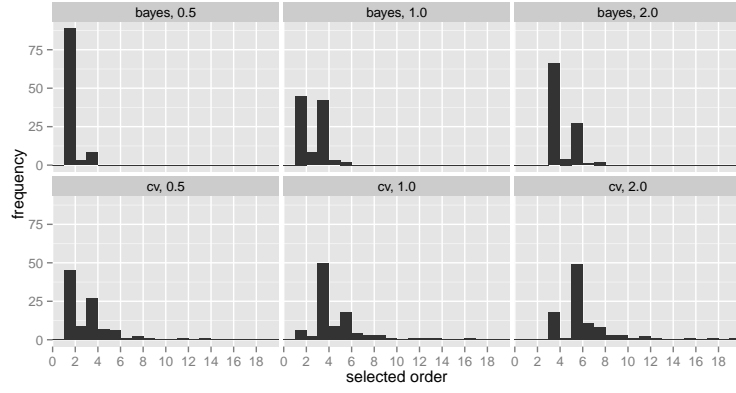
Figure 6: sup-norm for the difference between the true mean $\mu(x) = 5x(5x-0.2)(0.4x-1.8)(3x-1.8)(2x-1.8)$ and the fitted values of 1) the selected order \mathcal{N}^* polynomial with the proposed method (bayes), 2) the selected polynomial using cross-validation (cv), and 3) the full order \mathcal{N} model (full).



(a) $n = 100$

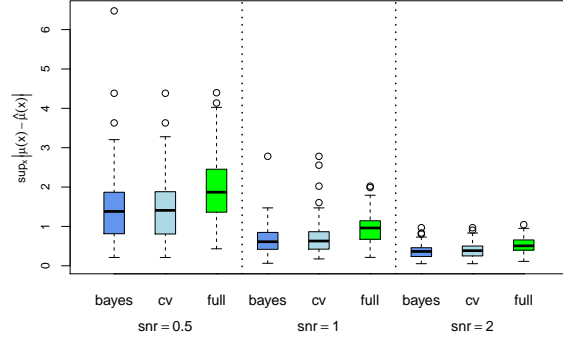


(b) $n = 200$

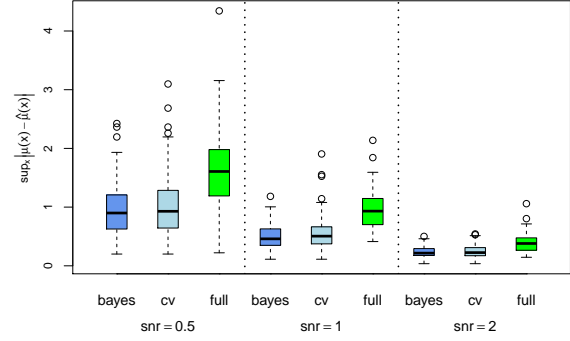


(c) $n = 500$

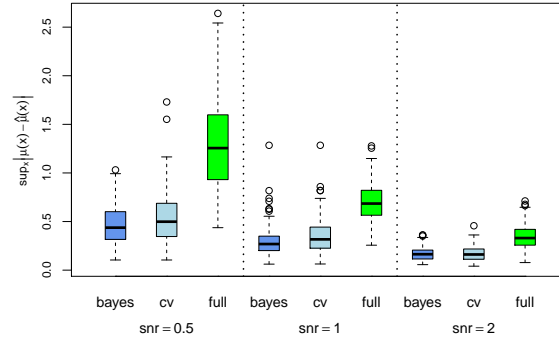
Figure 7: Frequencies for the order of smoothness selected in 100 simulated datasets with $SNR = 0.5, 1, 2$ for the proposed method (bayes) and cross-validation (cv) with mean function $\mu(x) = x\mathbf{I}_{\{x \in [-3, -1]\}} - \mathbf{I}_{\{x \in \{-1, 1\}\}} + (x - 2)\mathbf{I}_{\{x \in [1, 3]\}}$.



(a) $n = 100$



(b) $n = 200$



(c) $n = 500$

Figure 8: sup-norm for the difference between the true mean $\mu(x) = x\mathbf{I}_{\{x \in [-3, -1]\}} - \mathbf{I}_{\{x \in [-1, 1]\}} + (x - 2)\mathbf{I}_{\{x \in [1, 3]\}}$ and the fitted values of 1) the selected order \mathcal{N}^* polynomial with the proposed method (bayes), 2) the selected polynomial using cross-validation (cv), and 3) the full order \mathcal{N} model (full).

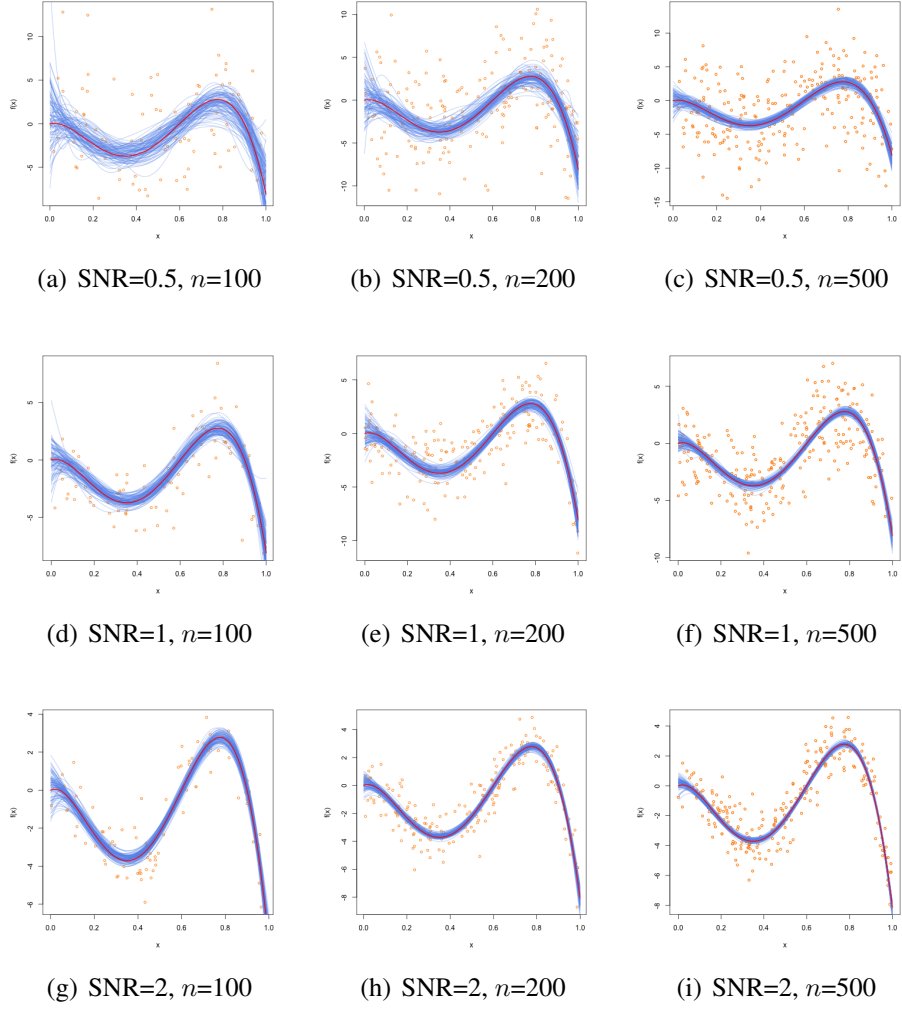


Figure 9: Model fit for the selected order of smoothness in 100 datasets using the proposed method with mean function $\mu(x) = 5x(5x - 0.2)(0.4x - 1.8)(3x - 1.8)(2x - 1.8)$

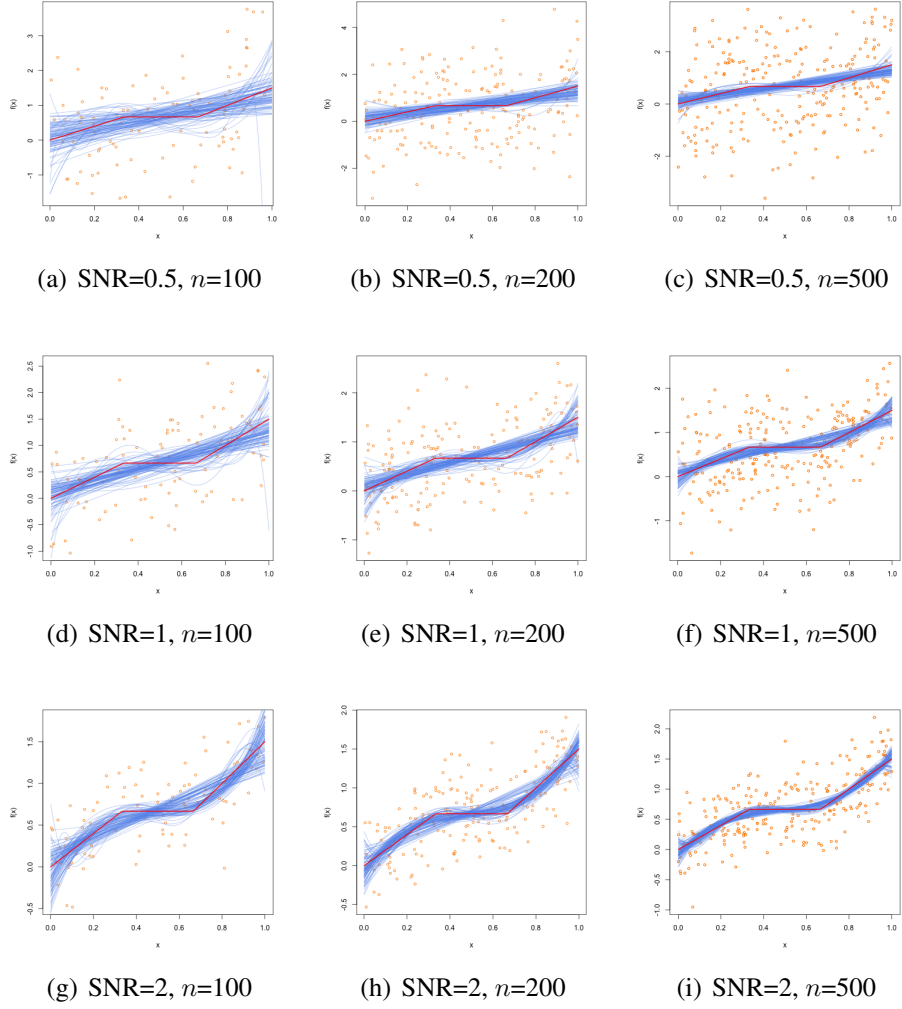


Figure 10: Model fit for the selected order of smoothness in 100 datasets using the proposed method with mean function $\mu(x) = x\mathbf{I}_{\{x \in [-3, -1]\}} - \mathbf{I}_{\{x \in \{-1, 1\}\}} + (x - 2)\mathbf{I}_{\{x \in [1, 3]\}}$